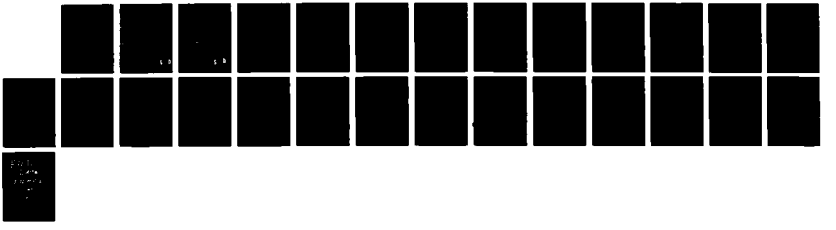
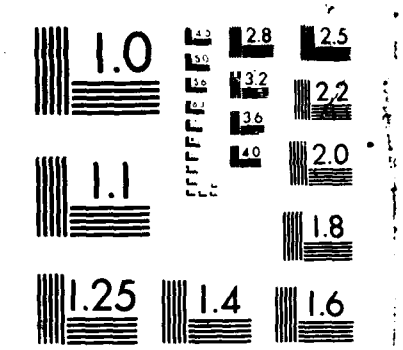
AD-A186 180 A SMOOTH NONPARAMETRIC QUANTILE ESTIMATOR FROM RIGHT-CENSORED DATA(U) SOUTH CAROLINA UNIV COLUMBIA DEPT OF STATISTICS W J PADGETT ET AL. MAY 87 TR-127 1/1

UNCLASSIFIED AFOSR-TR-87-1321 $AFOSR-84-0156 F/G 12/3 NL





DTIC FILE COPY

UNCLA
SECURIT

AD-A186 180 DOCUMENTATION PAGE

1a. REPC Uncla	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release, distribution unlimited
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) Stat. Tech. Rep. No. 127 (62.G05-20)	5. MONITORING ORGANIZATION REPORT NUMBER(S) AFOSR-TR- 87-1321
6a. NAME OF PERFORMING ORGANIZATION Department of Statistics	6b. OFFICE SYMBOL (If applicable)
7a. NAME OF MONITORING ORGANIZATION Air Force Office of Scientific Research	
6c. ADDRESS (City, State and ZIP Code) University of South Carolina Columbia, SC 29208	7b. ADDRESS (City, State and ZIP Code) Directorate of Mathematical & Information Sciences, Bolling AFB, DC 20332 Bldg 410
8a. NAME OF FUNDING/SPONSORING ORGANIZATION AFOSR	8b. OFFICE SYMBOL (If applicable) NM
9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER AFOSR-84-0156	
8c. ADDRESS (City, State and ZIP Code) Bolling AFB, DC 20332 Bldg 410	10. SOURCE OF FUNDING NOS. PROGRAM ELEMENT NO. 61102F; PROJECT NO. 2304; TASK NO. A5; WORK UNIT NO.

11. TITLE (Include Security Classification)
A Smooth Nonparametric Quantile Estimator from Right-Censored Data

12. PERSONAL AUTHOR(S)
W. J. Padgett and L. A. Thombs

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Yr., Mo., Day)	15. PAGE COUNT
Technical Minor	FROM _____ TO _____	May, 1987	22

16. SUPPLEMENTARY NOTATION
sub n

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB. GR.	Right-censoring; Percentiles; Asymptotic normality; Bootstrap method.

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Based on randomly right-censored data, a smooth nonparametric estimator of the quantile function of the lifetime distribution is studied. The estimator is defined to be the solution $x_n(p)$ to $F_n(x_n(p)) = 0$, where F_n is the distribution function corresponding to a kernel estimator of the lifetime density. The strong consistency and asymptotic normality of $x_n(p)$ are shown. Some simulation results comparing this estimator with the product-limit quantile estimator and a kernel-type estimator are presented. Data-based selection of the bandwidth required for computing F_n is investigated using bootstrap methods. Illustrative examples are given.

DTIC ELECTE OCT 0 6 1987 D

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
CLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☐ DTIC USERS ☐	Unclassified

22a. NAME OF RESPONSIBLE INDIVIDUAL	22b. TELEPHONE NUMBER (Include Area Code)	22c. OFFICE SYMBOL
Maj. Brian W. Woodruff	(202) 767-5027	NM

DD FORM 1473, 83 APR EDITION OF 1 JAN 73 IS OBSOLETE. UNCLASSIFIED

A SMOOTH NONPARAMETRIC QUANTILE ESTIMATOR
FROM RIGHT-CENSORED DATA

by

W. J. Padgett* and L. A. Thombs

University of South Carolina
Statistics Technical Report No. 127
62G05-20

DEPARTMENT OF STATISTICS

The University of South Carolina
Columbia, South Carolina 29208

A SMOOTH NONPARAMETRIC QUANTILE ESTIMATOR FROM RIGHT-CENSORED DATA

by

W. J. Padgett* and L. A. Thombs

University of South Carolina
Statistics Technical Report No. 127
62G05-20

May, 1987

Department of Statistics
University of South Carolina
Columbia, SC 29208

* Research partially supported by the U. S. Air Force Office of Scientific Research grant number AFOSR-84-0156 and the U. S. Army Research Office grant number MIPR ARO 139-85.

ABSTRACT

Based on randomly right-censored data, a smooth nonparametric estimator of the quantile function of the lifetime distribution is studied. The estimator is defined to be the solution $x_n(p)$ to $F_n(x_n(p))= p$, where F_n is the distribution function corresponding to a kernel estimator of the lifetime density. The strong consistency and asymptotic normality of $x_n(p)$ are shown. Some simulation results comparing this estimator with the product-limit quantile estimator and a kernel-type estimator are presented. Data-based selection of the bandwidth required for computing F_n is investigated using bootstrap methods. Illustrative examples are given.

Key words: Right-censoring; Percentiles; Asymptotic normality; Bootstrap method.

1. INTRODUCTION

Right-censored data arise frequently in industrial life testing experiments and medical studies. From such data it is important to be able to obtain good nonparametric estimates of various characteristics of the unknown underlying lifetime distribution. One characteristic of the lifetime distribution that is of interest is the quantile function. For example, in product development it is typical to estimate certain percentiles of the lifetime distribution of the product based on the right-censored observations from life tests. That is, estimates of possible guarantee times for the product are desired.

For a probability distribution function G, the quantile function is defined by $Q(p) \equiv G^{-1}(p) \equiv \inf \{t: G(t) \geq p\}$, $0 < p < 1$. For a random (uncensored) sample from G, several nonparametric estimates of Q(p) have been suggested. The sample quantile function, $G_n^{-1}(p) \equiv \inf \{x: G_n(x) \geq p\}$, has been studied, where $G_n(x)$ denotes the sample distribution function (see Csörgö , 1983, for example, for many of the known properties of G_n^{-1}). Another approach has been to solve $\tilde{G}_n(x_p) = p$ for x_p, where $\tilde{G}_n(x) = \int_{-\infty}^{x} g_n(t)dt$, with g_n being a kernel estimator of the density function of G (see Nadaraya, 1964). Recently, Yang (1985) studied a kernel-type estimator which smoothed the sample quantile function $G_n^{-1}(p)$.

For right-censored data, Sander (1975) proposed estimation of the quantile function by the product-limit (PL) estimator, defined by $\hat{Q}_n \equiv \hat{F}_n^{-1}$, where $\hat{F}_n$ denotes the PL estimator of the lifetime distribution (Kaplan and Meier, 1958; Efron, 1967). Sander (1975) and Cheng (1984) obtained some asymptotic properties of $\hat{Q}_n$, and Csörgö (1983) discussed strong approximation results. Padgett (1986) studied a kernel-type quantile estimator from right-censored observations, extending the complete sample results of Yang (1985). Lio, Padgett, and Yu (1986) and Lio and Padgett (1987) presented some asymptotic

properties of the kernel-type estimator, including asymptotic normality and mean square convergence. Also, the kernel-type estimator, Padgett and Thombs (1986) presented results of extensive simulations and investigated the use of bootstrap methods for bandwidth selection and confidence intervals.

In this paper, a smooth nonparametric estimator of the quantile function is studied which is defined as the solution $x_n(p)$ to $F_n(x_n(p)) = p$, where $F_n(x)$ is the distribution function corresponding to a kernel density estimator $f_n(x)$ of the lifetime density from right-censored data. The kernel density estimator proposed by Földes, Rejtö, and Winter (1980) and McNichols and Padgett (1986) is used here. The estimator $x_n(p)$ is intuitively more appealing than the kernel-type estimator of Padgett (1986) since it is a nondecreasing function of p. The kernel-type estimate can decrease for large p due to the scarcity of large uncensored observations in the sample.

The estimator $x_n(p)$ is defined in Section 2, and strong consistency and asymptotic normality are presented in Section 3. In Section 4, some simulation results are discussed comparing this estimator with the PL quantile estimator and the kernel-type estimator of Padgett (1986) with respect to estimated mean squared errors. Bootstrap methods for choosing a data-based bandwidth value for the kernel density estimate $f_n(x)$ are presented in Section 5. A confidence interval for the true lifetime distribution quantile based on the bootstrap percentile interval method is also given in that section.

2. NOTATION AND PRELIMINARIES

Let $X_1^o,\ldots,X_n^o$ denote the true survival times of n items or individuals that are censored on the right by a sequence $U_1,\ldots,U_n$, which in general may be either constants or random variables. The X_i^o's are nonnegative, independent, identically distributed random variables with common unknown continuous distribution function F_o, unknown quantile function $Q^o(p) \equiv F_o^{-1}(p) =$

$\inf\{t:F_o(t)\geq p\}$, $0<p<1$, and unknown density function f_o.

The observed right-censored data are denoted by the pairs (X_i,Δ_i), $i=1,\ldots,n$, where

$$X_i = \min\{X_i^o, U_i\},\ \Delta_i = \begin{cases} 1 \text{ if } X_i^o \leq U_i \\ 0 \text{ if } X_i^o > U_i \end{cases}.$$

Thus, it is known which observations are times of failure or death and which ones are censored or loss times. The nature of the censoring depends on the U_i's. (i) If $U_1,\ldots,U_n$ are fixed constants, the observations are time-truncated. If all U_i's are equal to the same constant, then the case of Type I censoring results. (ii) If all $U_i = X_{(r)}^o$, the rth order statistic of $X_1^o,\ldots,X_n^o$, then the situation is that of Type II censoring. (iii) If $U_1,\ldots,U_n$ constitute a random sample from a distribution H (usually unknown) and are independent of $X_1^o,\ldots,X_n^o$, then (X_i,Δ_i), $i=1,2,\ldots,n$, is called a randomly right-censored sample.

For the asymptotic results in Section 3 of this paper, the random censorship model (iii) is assumed. For this model the distribution function of each X_i is $F = 1 - (1-F_o)(1-H)$. This assumption is typically necessary for asymptotic results under censoring. For example, see Breslow and Crowley (1974), Földes, Rejtö and Winter (1980), Padgett (1986), and Lio, Padgett and Yu (1986).

A popular estimator of the survival function $1-F_o(t)$ from the censored sample (X_i,Δ_i), $i=1,\ldots,n$, is the product-limit (PL) estimator of Kaplan and Meier (1958). The PL estimator, which was shown to be "self-consistent" by Efron (1967), is defined as follows. Let (Z_i,Λ_i), $i=1,\ldots,n$, denote the ordered X_i's along with their corresponding Δ_i's. Values of the censored sample will be denoted by the corresponding lower case letters, (x_i,δ_i) and (z_i,λ_i), for the unordered and ordered sample, respectively. Then the PL estimator of $1-F_o(t)$ is

$$\hat{P}_n(t)= \begin{cases} 1, & 0\le t\le Z_1, \\ \prod_{i=1}^{k-1} \left(\frac{n-i}{n-i+1}\right)^{\Lambda_i}, & Z_{k-1}<t\le Z_k, k=2,\ldots,n \\ 0, & Z_n<t. \end{cases}$$

The PL estimator of $F_o(t)$ is denoted by $\hat{F}_n(t)=1-\hat{P}_n(t)$, and the size of the jump of $\hat{P}_n$ (or $\hat{F}_n$) at Z_j is denoted by s_j. Note that $s_j=0$ if and only if Z_j is censored for $j<n$, i.e. if and only if $\lambda_j=0$. Define $S_i = \sum_{j=1}^{i} s_j = \hat{F}_n(Z_{i+1})$, $i=1,\ldots,n-1$, and $S_n\equiv 1$.

A natural estimator of $Q^o(p)$ is the PL quantile function $\hat{Q}_n(p)=\inf\{t:\hat{F}_n(t)\ge p\}$ (see, for example, Sander (1975), Cheng (1984), and Csörgö (1983) for some of the properties of $\hat{Q}_n$). Since $\hat{Q}_n$ is a step function with jumps corresponding to the uncensored observations, it is desirable to obtain a smoothed estimator of Q^o. The kernel smoothed $\hat{Q}_n$, considered by Padgett (1986), Lio, Padgett and Yu (1986), Lio and Padgett (1987), and Padgett and Thombs (1986), is such an estimator, and is defined as follows: Let $\{h\equiv h_n\}$ be a "bandwidth" sequence of positive numbers so that $h_n\to 0$ as $n\to\infty$, and let K be a bounded probability density function which is zero outside a finite interval $(-c,c)$ and is symmetric about zero. (For asymptotic results, other conditions on h_n, K, and F_o are needed, but these are the only assumptions that will be made for purposes of definition.) Then for $0<p<1$, the kernel quantile function estimator is given by

$$\begin{aligned} Q_n(p) &= h^{-1}\int_0^1 \hat{Q}_n(t)K((t-p)/h)dt \\ &= h^{-1}\sum_{i=1}^{n} Z_i \int_{S_{i-1}}^{S_i} K((t-p)/h)dt, \end{aligned} \tag{2.1}$$

where $S_0\equiv 0$. An approximation to $Q_n(p)$ was given by Padgett (1986) as

$$\tilde{Q}_n(p)=h^{-1}\sum_{i=1}^{n} Z_i s_i K((S_i-p)/h). \tag{2.2}$$

Although neither estimator is difficult to compute (2.2) will be simpler for some kernel functions.

In this paper a different smooth nonparametric quantile estimator than (2.1) is studied. It is defined as the quantile function corresponding to the distribution function obtained from a kernel smoothed density estimate. To define this estimator from the ordered censored data (Z_i,Λ_i), i=1,...,n, consider the kernel density estimator of f_o written by McNichols and Padgett (1986) as

$$f_n(t) = h^{-1} \sum_{j=1}^{n} s_j K((t-Z_j)/h),\ t\geq 0,$$

where h and K are the bandwidth and kernel function defined earlier. The distribution function corresponding to the density f_n can be written as

$$F_n(x) = \int_{-\infty}^{x} f_n(t)dt = \int_0^{\infty} W((x-t)/h)d\hat{F}_n(t)$$

$$= \sum_{j=1}^{n} s_j\, W((x-Z_j)/h), \qquad (2.3)$$

where $W(t) = \int_{-\infty}^{t} K(u)du$ is the distribution function for the kernel K. Then the estimator $x_n(p)$ of the pth quantile, $Q^o(p)$, is defined to be the solution to the equation $F_n(x)=p$. This solution can be found iteratively by the Newton-Raphson method, for example, using a starting value such as the PL quantile, $\hat{Q}_n(p)$. In all computations reported in this paper, the iterations converged rapidly.

Although the estimate $x_n(p)$ must be obtained by an iterative procedure, whereas $Q_n(p)$ can be calculated directly, $x_n(p)$ is more appealing since it is always a nondecreasing function of p. The estimate $Q_n(p)$ can decrease for large p (see Figure 1 of Padgett, 1986). This is due to the scarcity of uncensored observations from the tail of F_o and can possibly be avoided by appropriately increasing the bandwidth h to compensate for fewer observations.

However, this would tend to oversmooth the estimate. A small computer simulation study to be discussed in Section 4 will indicate some further comparison of $x_n(p)$ and $Q_n(p)$ for small sample sizes. In fact, the simulation results indicate that, while both $x_n(p)$ and $Q_n(p)$ are better than $\hat{Q}_n(p)$ in the sense of smaller estimated mean squared errors for a range of bandwidth values, neither performs uniformly better than the other.

3. SOME ASYMPTOTIC RESULTS

In this section, the consistency and asymptotic normality of the estimator $x_n(p)$ will be presented assuming the random right-censorship model. The results will be stated, and their proofs will be outlined in the appendix. The consistency will be stated first.

For any distribution function G, define $T_G \equiv \sup \{t: G(t) < 1\}$.

Theorem 1. Suppose F_o is continuous and strictly increasing on $[0,\infty)$. If either (i) $H(T_{F_o}^-) < 1$, where t^- denotes limit from the left, or (ii) $T_{F_o} \leq T_H \leq \infty$, then $x_n(p) \to Q^o(p)$ almost surely as $n \to \infty$.

The conditions $H(T_{F_o}^-) < 1$ and $T_{F_o} \leq T_H \leq \infty$ guarantee a positive probability of observing uncensored data points from the entire support of the lifetime distribution F_o. The condition in (ii) allows both the lifetime distribution and censoring distribution to have the same support and to have support equal to the interval $[0,\infty)$. Hence, F_o and H can both be exponential, Weibull, or gamma distributions, for example.

As the next theorem states, $x_n(p)$ has the same limiting normal distribution as $Q_n(p)$ (Lio, Padgett and Yu, 1986). The proof, given in the Appendix, uses the concepts of Kiefer processes (see Csörgö, 1983).

Theorem 2. Let T satisfy $1-F(T)=[1-F_o(T)][1-H(T)]>0$. Assume that $Q^o(p)<T$, $f_o(1-H)$ is continuous and positive at $Q^o(p)$, the density function of H is continuous, and K is a continuous density defined on the finite interval $[-c,c]$. If $h\to 0$, $nh\to\infty$, and $\sqrt{n}h\to 0$ as $n\to\infty$, then $\sqrt{n}[x_n(p)-Q^o(p)]\to Z_p$ in distribution, where Z_p is a normally distributed random variable with mean zero and variance

$$\sigma_p^2 = (1-p)^2[f_o(Q^o(p))]^2 \int_0^{Q^o(p)} [1-F(u)]^{-2}\, d\tilde{F}_o(u),$$

with $\tilde{F}_o(u) = P(X\le u, \Delta=1)$ denoting the subdistribution function of the uncensored observations.

An example of a bandwidth sequence $\{h_n\}$ satisfying the conditions of Theorem 2 is $h_n=cn^{-d}$, $\frac{1}{2}<d<1$.

4. SOME SIMULATION RESULTS

A small Monte Carlo simulation was performed to obtain an indication of the performance of $x_n(p)$ compared with the kernel-type estimator, $Q_n(p)$, for small sample sizes. For these simulations the triangular density on $[-1,1]$ was used for K, $K(u)=1-|u|$, $|u|\le 1$. The censoring distribution H was taken to be the exponential distribution with mean β^{-1} and the lifetime distributions used were Weibull with shape parameter α and scale parameter equal to one, that is,

$F_o(x) = 1-e^{-x^\alpha}$ (α=0.5,1, and 2).

The bandwidth values of h=0.01 (0.04) 0.61 were used for quantiles at p= 0.10, 0.25, 0.50, 0.75. Sample sizes of n= 30,60,100 were studied.

In each case simulated (i.e. each distribution, bandwidth, p, and sample size combination), 300 censored samples were generated using the random number generators in the International Mathematical and Statistical Library (IMSL,

1985) on a VAX 11/8300 computer. From the 300 samples the estimated mean squared errors (Average Squared Error≡ASE) of the estimators $x_n(p)$, $Q_n(p)$ and $\hat{Q}_n(p)$ were computed, and the ratios of these ASE's, $ASE[\hat{Q}_n(p)]/ASE[x_n(p)]$ and $ASE[Q_n(p)]/ASE[x_n(p)]$, were calculated.

Some of the results of the simulations are given in Tables 1-3. In all cases for each p, except for small p for the Weibull lifetime distribution with $\alpha=0.5$, there is a range of bandwidth values for which $x_n(p)$ has smaller ASE than that of the PL quantile estimator, $\hat{Q}_n(p)$. Also, in many of the simulated cases, there is a range of h values for which the ASE of $x_n(p)$ is less than that of the kernel estimator, $Q_n(p)$. This is the case for the exponential lifetime distribution for all values of p simulated. However, neither $x_n(p)$ nor $Q_n(p)$ is uniformly better than the other over all values of p and bandwidths used in the simulations.

In the next section the bootstrap will be used to determine, based on the given censored sample, the "best" value of h to use (in the sense of the smallest bootstrap MSE) in calculating $x_n(p)$ as p varies. Bootstrap confidence bounds for the true quantile $Q^o(p)$ will also be discussed, and an example using switch failure data adapted from Nair (1984) will be presented.

5. BOOTSTRAP METHODS: BANDWIDTH SELECTION AND CONFIDENCE INTERVALS

Since the estimator $x_n(p)$ is implicitly defined as the solution to $F_n(x_n(p)) = p$, where $F_n(x) = \int_{-\infty}^{x} f_n(t)dt$, it depends on a bandwidth value h_n used in the kernel smoothed density estimate $f_n(t)$. Thus, in practice h_n must be chosen before the estimator $x_n(p)$ can be computed. A natural question to ask is: "Which bandwidth value yields the 'best' estimate $x_n(p)$ of $Q^o(p)$, in the sense of minimum mean square error (MSE)?" Due to the censoring, exact or even asymptotic expressions for $MSE(x_n(p))$ are difficult to derive. In this section we propose a method of selecting bandwidths based on minimizing the bootstrap estimate of $MSE(x_n(p))$.

TABLE 1. Ratios of Average Squared Errors (300 samples)
Exponential (λ=1) Life Distribution and
Exponential (β=3/7) Censoring Distribution (30% censoring)

n=30

p \ h		.01	.05	.09	.13	.17	.21	.25	.29	.33	.37	.45	.57
.10	a	.127	1.016	1.028	.965	1.035	1.044	1.024	1.206	1.126	.867	.794	.611
	b	.123	1.082	1.046	.940	.909	.880	.951	1.124	1.152	1.370	1.923	1.742
.25	a	.226	.026	1.027	1.121	1.183	1.105	1.192	1.119	1.332	1.197	1.345	1.238
	b	.216	.235	.958	.938	.995	.988	.976	1.040	1.020	1.093	1.295	1.689
.50	a	1.008	1.048	1.061	1.044	1.055	1.092	1.093	1.136	1.134	1.174	1.156	1.241
	b	.972	.957	.932	.930	.960	.976	1.032	1.088	1.016	1.052	1.538	1.645
.75	a	1.005	1.021	1.022	1.033	1.068	1.086	1.112	1.074	1.191	1.149	1.177	1.205
	b	.865	.938	1.058	.941	.839	.964	.786	.864	.911	.595	.515	.536

n=60

p \ h		.01	.05	.09	.13	.17	.21	.25	.29	.33	.37	.45	.57
.10	a	1.018	1.116	1.136	1.146	1.366	1.204	1.163	1.140	1.060	1.030	.745	.266
	b	1.018	.978	.958	.943	.938	.982	1.022	1.260	1.287	1.484	1.725	1.604
.25	a	1.006	1.033	1.102	1.111	1.178	1.198	1.274	1.244	1.276	1.290	1.296	1.248
	b	.979	1.001	.977	.992	.999	.999	.969	1.030	1.082	1.102	1.325	2.660
.50	a	1.007	1.033	1.040	1.093	1.134	1.119	1.155	1.147	1.129	1.234	1.205	1.285
	b	.974	.976	.956	.964	.945	.994	.991	1.046	1.040	1.205	1.386	2.659
.75	a	1.003	1.027	1.029	1.063	1.034	1.069	1.100	1.133	1.174	1.165	1.193	1.203
	b	.959	.870	.962	.953	1.080	1.002	.962	1.256	1.076	.948	.817	.859

n=100

p \ h		.01	.05	.09	.13	.17	.21	.25	.29	.33	.37	.45	.57
.10	a	1.024	1.125	1.216	1.291	1.265	1.267	1.275	1.326	1.177	.732	.559	.227
	b	.980	.983	.958	.951	.923	.998	1.084	1.666	1.875	1.621	1.858	1.682
.25	a	1.017	1.033	1.041	1.228	1.184	1.206	1.220	1.212	1.317	1.294	1.253	1.189
	b	.988	.991	.988	.977	.983	1.013	1.022	1.069	1.121	1.264	1.524	2.897
.50	a	1.009	1.014	1.035	1.059	1.107	1.173	1.132	1.064	1.218	1.168	1.238	1.243
	b	.985	.996	.938	.955	.971	.971	1.018	1.079	1.294	1.253	1.924	3.112
.75	a	1.007	1.014	1.038	1.089	1.072	1.078	1.088	1.093	1.089	1.209	1.248	1.202
	b	.954	.943	.906	.879	.934	1.202	1.459	1.617	1.153	.925	.816	1.419

a. $ASE(\hat{Q}_n)/ASE(x_n)$

b. $ASE(Q_n)/ASE(x_n)$

TABLE 2. Ratios of Average Squared Errors (300 samples)
Weibull (α=0.5, λ=1) Life Distribution and
Exponential (β=3/7) Censoring Distribution (69.7% censoring)

n=30

p \ h		.01	.05	.09	.13	.17	.21	.25	.29	.33	.37	.45	.57
.10	a	.795	.636	.443	.443	.507	.486	.280	.231	.114	.074	.034	.028
	b	1.229	1.049	.946	1.006	.968	.975	.821	1.009	.721	.725	.925	1.087
.25	a	1.021	1.120	1.122	1.127	1.196	1.129	1.024	1.091	.931	.787	.818	.674
	b	1.010	.963	1.025	1.099	1.163	1.220	1.241	1.523	1.586	2.081	3.183	6.655
.50	a	1.006	1.019	1.037	1.050	1.064	1.076	1.128	1.121	1.152	1.168	1.143	1.174
	b	.966	.977	1.009	.993	1.123	1.226	1.436	1.443	1.748	2.131	3.449	4.194
.75	a	1.002	1.010	1.018	1.022	1.031	1.036	1.052	1.051	1.071	1.075	1.086	1.133
	b	.971	.866	.898	.875	.778	.647	.608	.596	.579	.452	.347	.265

n=60

p \ h		.01	.05	.09	.13	.17	.21	.25	.29	.33	.37	.45	.57
.10	a	.827	.650	.652	.552	.391	.317	.109	.073	.051	.045	.032	.060
	b	1.070	.895	.860	.889	.781	.913	.534	.558	.459	.552	.624	.623
.25	a	1.007	1.029	1.125	1.047	.959	.831	.851	.687	.614	.588	.509	.401
	b	1.020	1.070	1.046	1.084	1.174	1.116	1.234	1.360	1.741	1.947	3.126	9.496
.50	a	1.003	1.010	1.024	1.063	1.052	1.136	1.065	1.175	1.086	1.167	1.137	1.119
	b	.991	.988	.925	1.033	1.127	1.341	1.762	1.902	2.023	3.505	4.118	11.417
.75	a	1.003	1.010	1.021	1.030	1.032	1.054	1.040	1.065	1.081	1.073	1.117	1.141
	b	.865	.961	.782	.759	.819	.794	.833	.756	.779	.436	.396	.234

a. $ASE(\hat{Q}_n)/ASE(x_n)$

b. $ASE(Q_n)/ASE(x_n)$

TABLE 3. Ratios of Average Squared Errors (300 samples)
Weibull (α=2, λ=1) Life Distribution and
Exponential (β=3/7) Censoring Distribution (27.6% censoring)

n=30

p		h	.01	.05	.09	.13	.17	.21	.25	.29	.33	.37	.45	.57
.10	a		.939	1.081	.981	1.220	1.226	1.318	1.386	1.565	1.531	1.348	1.417	1.052
	b		.872	.904	.720	.801	.745	.664	.685	.593	.564	.544	.451	.306
.25	a		.708	.429	1.083	1.125	1.180	1.156	1.312	1.287	1.393	1.290	1.493	1.464
	b		.678	.404	.956	.966	.950	.946	.899	.883	.866	.855	.780	.702
.50	a		1.012	1.015	1.114	1.118	1.104	1.161	1.161	1.276	1.316	1.326	1.243	1.302
	b		.991	.990	.978	.947	.957	.976	.958	.963	.940	.928	.987	.906
.75	a		1.012	1.023	1.065	1.091	1.109	1.166	1.168	1.220	1.292	1.415	1.532	1.345
	b		.967	1.001	.958	.937	.872	.978	.924	.863	.913	.820	1.276	2.537

n=60

p		h	.01	.05	.09	.13	.17	.21	.25	.29	.33	.37	.45	.57
.10	a		1.030	1.125	1.210	1.267	1.454	1.373	1.400	1.394	1.394	1.374	1.123	.536
	b		.981	.930	.876	.821	.791	.716	.776	.738	.650	.542	.371	.163
.25	a		1.009	1.043	1.104	1.097	1.166	1.257	1.289	1.294	1.209	1.365	1.340	1.220
	b		.997	.996	.965	.984	.939	.927	.905	.887	.902	.890	.841	.749
.50	a		1.000	1.002	1.076	1.088	1.149	1.207	1.189	1.203	1.220	1.293	1.303	1.439
	b		.993	.984	.961	.972	.958	.979	.977	.976	.962	.963	.948	.927
.75	a		1.000	1.027	1.095	1.128	1.111	1.151	1.144	1.227	1.217	1.247	1.350	1.165
	b		1.005	.965	.959	.957	.954	.984	1.025	.951	.914	.929	2.011	4.094

a. $ASE(\hat{Q}_n)/ASE(x_n)$

b. $ASE(Q_n)/ASE(x_n)$

Marron and Padgett (1987) have determined an asymptotically optimal bandwidth for the density estimate $f_n(t)$ which minimizes the integrated squared error (ISE) of f_n. Since their bandwidth value is based on minimizing the (asymptotic) global ISE of f_n, it does not work well in the setting of quantile estimation for small samples. Bandwidth values for the quantile estimates considered here depend on p, and results regarding bandwidth selection for $f_n(t)$ with respect to the asymptotic *minimum* ISE criteria do not carry over to the estimator $x_n(p)$.

Recently, the scope of the bootstrap has been extended from the iid case to include more complex data structures such as censored data and correlated data (see Efron and Tibshirani, 1986). By creating bootstrap replicates which are intended to "mimic" the statistical properties of the sample (and thus the population) one can learn about the sampling distribution of a statistic, regardless of its complicated form. In this paper the nonparametric bootstrap for censored data is used to investigate the $MSE(x_n(p))$ as a function of bandwidth.

Recall that (X_i,Δ_i), i=1,...,n, denotes the observed censored sample. Unlike the iid case, there is not a well-defined method for obtaining a bootstrap replicate (X_i^*,Δ_i^*), i=1,...,n. There have been several proposed methods for resampling censored data. Reid (1981) proposed resampling from the Kaplan-Meier estimator $\hat{F}_n$ of F_o, which results in bootstrap samples that contain only uncensored observations. In Efron's (1981) plan, one simply takes a random sample with replacement from $(X_1,\Delta_1),\ldots,(X_n,\Delta_n)$. While Reid's approach is analogous to resampling in the uncensored case, it is not clear what is being estimated by the bootstrap since no censored observations are present in any of the bootstrap replicates. Akritas (1986) studied the asymptotic properties of Efron's and Reid's procedures for bootstrapping censored survival data. He showed that Efron's approach yields asymptotically correct confidence bands for

F_o based on the Kaplan-Meier estimate $\hat{F}_n$, while Reid's does not. Since $x_n(p)$ involves $\hat{F}_n$, we have adopted Efron's approach of drawing at random, with replacement, from the n data values to get (X_i^*, Δ_i^*), i=1,...,n.

For fixed p and h, we define the bootstrap estimate of MSE($x_n(p)$ as follows: For each bootstrap replicate, the quantile estimate $x_n^*(p)$ is calculated. As is usually the case, a large number, B, of bootstrap samples and the corresponding estimates $x_n^*(p)$ are obtained. The bootstrap estimate of variance is given by

$$\text{Var}^*(x_n(p)) = \frac{1}{B-1} \sum_{i=1}^{B} \left[x_n^{*i}(p) - \bar{x}_n^*(p)\right]^2, \tag{5.1}$$

where $x_n^{*i}(p)$ denotes the estimate $x_n^*(p)$ calculated from the ith bootstrap replicate and $\bar{x}_n^*(p) = \sum_{i=1}^{B} x_n^{*i}(p)/B$. The bias estimate is

$$\text{Bias}^*(x_n(p)) = \bar{x}_n^*(p) - x_n(p), \tag{5.2}$$

where $x_n(p)$ is the estimate calculated from the original data. Then for a bandwidth value h and fixed p, the bootstrap estimate of MSE($x_n(p)$) is

$$\text{MSE}^*(h) = \text{Var}^*(x_n(p)) + [\text{bias}^*(x_n(p))]^2.$$

The value h^* which yields minimum $\text{MSE}^*(h)$ is the bandwidth selected to calculate $x_n(p)$.

Once the bandwidth h^* has been selected, the set of B values $x_n^{*1}(p),\ldots,$ $x_n^{*B}(p)$ corresponding to h^* can be used to construct a confidence interval for $Q^o(p)$. Define the bootstrap cumulative distribution function of $x_n(p)$ by

$$G^*(y) = \{\text{number of values } x_n^{*i}(p) \leq y\} / B. \tag{5.3}$$

Then the endpoints of the interval are quantiles of G^*; that is, a 100(1-α)% confidence interval for $Q^o(p)$ is given by

$$[G^{*-1}(\alpha/2),\ G^{*-1}(1-\alpha/2)]. \tag{5.4}$$

Note that this is an application of Efron's (1980) percentile interval method. A refinement of this method, called the "bias-corrected acceleration constant percentile interval," has recently been proposed by Efron (1987) but was not considered here.

To illustrate the performance of bootstrap bandwidth selection and confidence intervals, a sample of size n=60 was generated using an exponential (mean=1) lifetime distribution and exponential (mean=1) censoring distribution. The triangular density on [-1,1] was used as the kernel function K(u) for f_n. Bandwidth values h=.02(.02).60 were considered and quantile estimates for p=.025(.025).975, as well as p=.01 and .99, were studied. For the calculation of bootstrap MSE, B=300 was used. Some results are given in Table 4 and Figure 1. Note that the values of h^* are small for small p, increase for p up to about 0.75, and then tend to decrease for larger p. For p=.10, Figure 2 shows a graph of bootstrap $MSE^*(h)$ vs. h. The general quadratic shape of the MSE curve is obvious. Table 4 indicates that the estimator $x_n(p)$ is often very close to the product-limit estimator, but when the two differ, $x_n(p)$ is closer to the true quantile. In Figure 1, the close agreement of $x_n(p)$ to $Q^o(p)$ for all p is illustrated. Note that the confidence bands become wider as p increases. This was the case in all the simulations and is probably a result of the random right censoring present in the data. In order to better estimate the sampling distribution of $x_n(p)$, B=1000 was used in obtaining confidence bounds.

Next, as an application of the bootstrap bandwidth selection procedure, we consider mechanical switch failure data adapted from Nair (1984) (see Table 5). The triangular kernel was used and estimators were obtained for p=.05(.05).95. Table 6 gives the estimates, bootstrap bandwidths, and MSE^* values for some values of p. The quantile estimate and confidence bands are plotted in Figure 3.

TABLE 4. Bootstrap Bandwidth Selection for $x_n(p)$

Exponential Lifetime Distribution (mean=1) and
Exponential Censoring Distribution (mean=1).

n=60

p	h^*	$x_n(p)$	$\hat{Q}_n(p)$	$Q^0(p)$
.01	.02	-.001	.001	.010
.10	.26	.101	.053	.105
.25	.06	.310	.311	.288
.50	.58	.757	.755	.693
.75	.54	1.523	1.521	1.386
.90	.58	2.344	2.106	2.303
.99	.40	4.627	4.491	4.605

TABLE 5. Failure Times (in Millions of Operations) of Switches

Z_i	Λ_i	Z_i	Λ_i	Z_i	Λ_i	Z_i	Λ_i
1.151	0	1.667	1	2.119	0	2.547	1
1.170	0	1.695	1	2.135	1	2.548	1
1.248	0	1.710	1	2.197	1	2.738	0
1.331	0	1.955	0	2.199	0	2.794	1
1.381	0	1.965	1	2.227	1	2.883	0
1.499	1	2.012	0	2.250	0	2.883	0
1.508	0	2.051	0	2.254	1	2.910	1
1.543	0	2.076	0	2.261	0	3.015	1
1.577	0	2.109	1	2.349	0	3.017	1
1.584	0	2.116	0	2.369	1	3.793	0

TABLE 7. Quantile Estimates for Switch Data

p	h^*	$x_n(p)$	MSE^*
.05	.10	1.655	.01027
.25	.24	2.165	.01323
.50	.60	2.581	.01923
.75	.44	3.015	.03188
.95	.60	3.785	.10917

FIGURE 1. $x_n(p)$ and Confidence Intervals for Simulated Data.

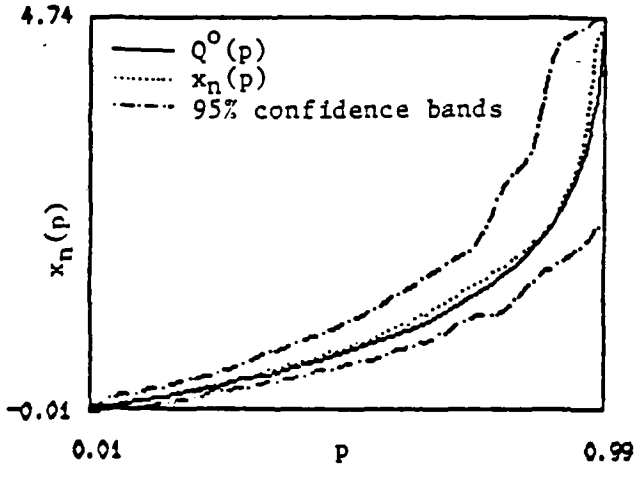


FIGURE 2. Bootstrap MSE vs. h_n (p=.10).

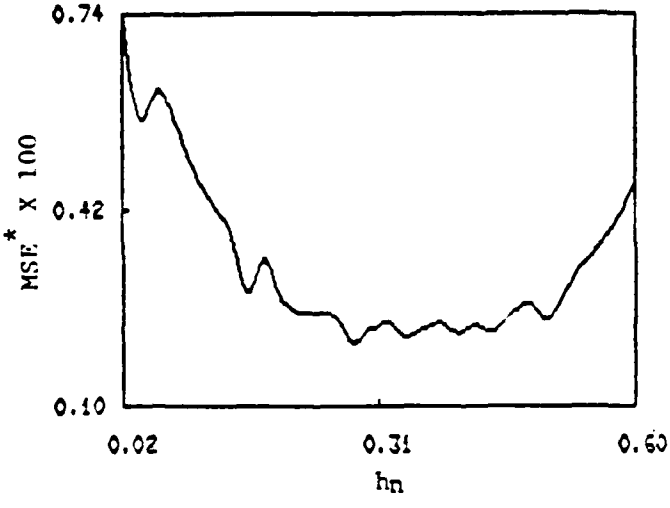
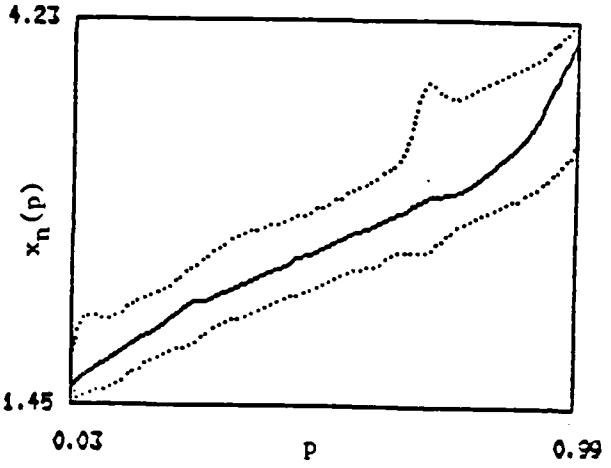


FIGURE 3. $x_n(p)$ for Switch Failure Data.



APPENDIX: PROOFS OF THEOREMS

Denote $Q^o(p) \equiv \xi_p^o$.

Proof of Theorem 1(i). For fixed $0<p<1$, write

$$\begin{aligned}|F_o(x_n(p)) - F_o(\xi_p^o)| &= |F_o(x_n(p))-p| \\ &= |F_o(x_n(p))-F_n(x_n(p))| \\ &\leq \sup_x |F_o(x)-F_n(x)|. \end{aligned} \tag{A.1}$$

The right-hand side of (A.1) converges to zero almost surely by Theorem 5.3 of Földes, Rejtö and Winter (1980). Since F_o^{-1} is continuous, then $x_n(p) \to \xi_p^o$ almost surely. ///

To prove part (ii) of Theorem 1, the following lemma is needed. Let $K_n(\cdot) \equiv W(\cdot/h_n)$ and $\bar{K}_n(y;x) \equiv 1-K_n(x-y)$.

Lemma. Let F_o be continuous and $T_{F_o} \leq T_H \leq \infty$. Then $\sup_{-\infty<x\leq T_{F_o}} |F_n(x)-F_o(x)| \to 0$ almost surely.

Proof. By Corollary 2(ii) of Csörgö and Horváth (1983), $\sup_{-\infty<x\leq T_{F_o}} |\hat{F}_n(x)-F_o(x)| \to 0$ almost surely. From Lemma 5.2 of Földes, Rejtö and Winter (1980) and by definition of T_{F_o}, letting $\bar{F}_n(x) = \int F_o(y)d\bar{K}_n(y;x)$,

$$\begin{aligned}|F_n(x)-\bar{F}_n(x)| &= |\int \hat{F}_n(y)d\bar{K}_n(y;x)-\bar{F}_n(x)| \\ &= \left| \int_{-\infty}^{T_{F_o}} [\hat{F}_n(y)-F_o(y)]d\bar{K}_n(y;x) \right| \\ &\leq \sup_{-\infty<x\leq T_{F_o}} |\hat{F}_n(x)-F_o(x)| \int_{-\infty}^{T_{F_o}} d\bar{K}_n(y;x) \\ &\leq \sup_{-\infty<x\leq T_{F_o}} |\hat{F}_n(x)-F_o(x)|. \end{aligned}$$

Thus, $\sup_{-\infty<x\leq T_{F_o}} |F_n(x)-\bar{F}_n(x)| \to 0$ almost surely.

Now, given $\varepsilon>0$, let $\delta>0$ be such that $|y|\leq\delta$ implies $\sup_x |F_o(x-y)-F_o(x)|<\varepsilon$.

Then

$$|\bar{F}_n(x)-F_o(x)| = \left| \int F_o(x-y)dK_n(y) - \int F_o(x)dK_n(y) \right|$$

$$\leq \varepsilon + \int_{|y|\leq\delta} dK_n(y). \qquad (A.2)$$

But as $n\to\infty$ the last integral in (A.2) approaches zero, so $\sup_x |\bar{F}_n(x)-F_o(x)| \to 0$.

The result of the lemma now follows by the triangle inequality. ///

Proof of Theorem 1(ii). The result follows from the Lemma, the continuity of F_o^{-1}, and writing

$$|F_o(x_n(p))-F_o(\xi_p^o)| = |F_o(x_n(p))-p|$$

$$\leq |F_n(x_n(p))-F_o(x_n(p))|$$

$$\leq \sup_{-\infty<x\leq T_{F_o}} |F_n(x)-F_o(x)|. \quad ///$$

Proof of Theorem 2. Approximating $F_n(x_n(p))=p$ by its two-term Taylor expansion about ξ_p^o, write

$$\sqrt{n}\, f_o(\xi_p^o)[x_n(p)-\xi_p^o] = -\sqrt{n}\, [F_n(\xi_p^o)-F_o(\xi_p^o)]\, \frac{f_o(\xi_p^o)}{f_n(\xi)}, \qquad (A.3)$$

where ξ is some (random) point between $x_n(p)$ and ξ_p^o.

Now, $|f_n(\xi)-f_o(\xi_p^o)| \leq |f_n(\xi)-f_o(\xi)| + |f_o(\xi)-f_o(\xi_p^o)|$. Under the assumptions of the theorem, by Corollary 2 of Mielniczuk (1986),

$$\sup_{0\leq x\leq T} |f_n(x)-f_o(x)| \to 0 \text{ almost surely.}$$

From Theorem 1, $x_n(p)\to\xi_p^o$ almost surely, so $\xi\to\xi_p^o$ almost surely. Hence, by the continuity of f_o, $|f_n(\xi)-f_o(\xi_p^o)|\to 0$ almost surely. Therefore, (A.3) has the same limiting distribution as $\sqrt{n}\,[F_o(\xi_p^o)-F_n(\xi_p^o)]$.

Now, consider $\sqrt{n}[F_n(\xi_p^o)-F_o(\xi_p^o)] = I + II$, where $I = \sqrt{n}[F_n(\xi_p^o)-\hat{F}_n(\xi_p^o)]$ and $II = \sqrt{n}[\hat{F}_n(\xi_p^o)-F_o(\xi_p^o)]$. Next, write

$$I = \sqrt{n}\ \left[\int_0^{\xi_p^o} \int_0^{ch+u} h^{-1}K((u-x)/h)d\hat{F}_n(x)du - \hat{F}_n(\xi_p^o) \right]$$

$$= \sqrt{n}\ \left[\int_0^{ch+\xi_p^o} \int_{x-ch}^{x+ch} h^{-1}K((u-x)/h)\ du\ d\hat{F}_n(x) - \int_0^{\xi_p^o} d\hat{F}_n(x) \right].$$

Since $\int h^{-1}K((u-x)/h)du \leq 1$, for n sufficiently large,

$$0 < \int_{x-ch}^{x+ch} h^{-1}\ K((u-x)/h)du < 1, \text{ so}$$

$$I \leq \sqrt{n}\ \int_{\xi_p^o}^{ch+\xi_p^o} d\hat{F}_n(x) = \sqrt{n}[\hat{F}_n(\xi_p^o + ch) - \hat{F}_n(\xi_p^o)]. \qquad (A.4)$$

Next, letting K(t,s) denote a Kiefer process and $\beta_n(t) \equiv \sqrt{n}[\hat{F}_n(t)-F_o(t)]$ be the PL process (see Csörgö 1983, for these definitions), the right side of (A.4) can be written as

$$\begin{aligned}
&[\beta_n(\xi_p^o+ch)-n^{-\frac{1}{2}}K(\xi_p^o+ch,n)] \\
&\quad - [\beta_n(\xi_p^o)-n^{-\frac{1}{2}}K(\xi_p^o,n)] \\
&\quad + n^{-\frac{1}{2}}[K(\xi_p^o+ch,n)-K(\xi_p^o,n)] \\
&\quad + n^{\frac{1}{2}}[F_o(\xi_p^o+ch)-F_o(\xi_p^o)]. \qquad (A.5)
\end{aligned}$$

From Theorem A of Aly, Csörgö, and Horváth (1985), under the assumptions of the theorem, the first two terms of (A.5) converge to zero as $n\to\infty$. Also, by a proof similar to that of Lemma 1 of Lio, Padgett, and Yu (1986), the third term of (A.5) converges to zero in probability as $n\to\infty$. For the last term of (A.5), since $f_o(\xi_p^o)>0$ by assumption,

$$ch\ \sqrt{n}\ \left[\frac{F_o(\xi_p^o+ch)-F_o(\xi_p^o)}{ch} \right] \to 0$$

as $n\to\infty$ since $\sqrt{n}h\to 0$ by the conditions of the theorem. Therefore, $I\to 0$ in probability as $n\to\infty$. Hence, $\sqrt{n}[F_n(\xi_p^o)-F_o(\xi_p^o)]$ has the same well-known limiting distribution as II, which is the normal distribution with mean zero and variance $(1-p)^2 \int_0^{\xi_p^o} [1-F(u)]^{-2}\ d\tilde{F}_o(u)$, completing the proof.///

REFERENCES

Akritas, M. G. (1986), "Bootstrapping the Kaplan-Meier Estimator," Journal of the American Statistical Association, 81, 1032-1038.

Aly, E.-E.A.A., Csörgö, M. and Horváth, L. (1985), "Strong Approximations of the Quantile Process of the Product-Limit Estimator," Journal of Multivariate Analysis, 16, 185-120.

Breslow, N. and Crowley, J. (1974), "A Large Sample Study of the Life Table and Product-Limit Estimator under Random Censorship," Annals of Statistics, 2, 437-453.

Cheng, K. F. (1984), "On Almost Sure Representations for Quantiles of the Product-Limit Estimator with Applications," Sankhya, Ser. A, 46, 426-443.

Csörgö, M. (1983), Quantile Processes with Statistical Applications (CBMS-NSF Regional Conference Series in Applied Mathematics, No. 42), Philadelphia: Society for Industrial and Applied Mathematics.

Csörgö, S. and Horváth, L. (1983), "The Rate of Strong Uniform Consistency for the Product-Limit Estimator," Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete, 62, 411-426.

Efron, B. (1967), "The Two-Sample Problem with Censored Data," in Proceedings of the Fifth Berkeley Symposium (Vol. 4), Berkeley, CA: University of California Press, pp. 831-853.

Efron, B. (1980), The Jackknife, the Bootstrap, and Other Resampling Plans (CBMS-NSF Regional Conference Series in Applied Mathematics, No. 38), Philadelphia: Society for Industrial and Applied Mathematics.

Efron, B. (1981), "Censored Data and the Bootstrap," Journal of the American Statistical Association, 76, 312-319.

Efron, B. (1987), "Better Bootstrap Confidence Intervals," Journal of the American Statistical Association, 82, 171-185.

Efron, B. and Tibshirani, R. (1986), "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy," Statistical Science, 1, 54-75.

Földes, A., Rejtö, L. and Winter, B. B. (1980), "Strong Consistency Properties of Nonparametric Estimators for Randomly Censored Data, I: The Product-Limit Estimator," Periodica Mathematica Hungarica, 11, 233-250.

International Mathematical and Statistical Libraries, Inc. (1985), IMSL, Houston.

Kaplan, E. L. and Meier, P. (1958), "Nonparametric Estimation from Incomplete Observations," Journal of the American Statistical Association, 53, 457-481.

Lio, Y. L. and Padgett, W. J. (1987), "Some Convergence Results for Kernel-type Quantile Estimators under Censoring," Statistics and Probability Letters, 5, 5-14.

Lio, Y. L., Padgett, W. J., and Yu, K. F. (1986), "On the Asymptotic Properties of a Kernel-Type Quantile Estimator from Censored Samples," Journal of Statistical Planning and Inference, 14, 169-177.

Marron, J. S. and Padgett, W. J. (1987), "Asymptotically Optimal Bandwidth Selection for Kernel Density Estimators from Randomly Right-Censored Samples," Annals of Statistics (to appear).

McNichols, D. T. and Padgett, W. J. (1986), "Mean and Variance of a Kernel Density Estimator under the Koziol-Green Model of Random Censorship," Sankhya, Series A, 48, 150-168.

Mielniczuk, J. (1986), "Some Asymptotic Properties of Kernel Estimators of a Density Function in Case of Censored Data," Annals of Statistics, 14, 766-773.

Nadaraya, E. A. (1964), "Some New Estimates for Distribution Functions," Theory of Probability and Its Applications, 9, 497-500.

Nair, V. N. (1984), "Confidence Bands for Survival Functions with Censored Data: A Comparative Study," Technometrics, 26, 265-275.

Padgett, W. J. (1986), "A Kernel-Type Estimator of a Quantile Function from Right-Censored Data," Journal of the American Statistical Association, 81, 215-222.

Padgett, W. J. and Thombs, L. A. (1986), "Smooth Nonparametric Quantile Estimation Under Censoring: Simulations and Bootstrap Methods," Communications in Statistics, Simulation and Computation, 15, 1003-1025.

Reid, N. (1981), "Estimating the Median Survival Time," Biometrika, 68, 601-608.

Sander, J. (1975), "The Weak Convergence of Quantiles of the Product-Limit Estimator," Technical Report Number 5, Stanford University, Department of Statistics.

Yang, S. S. (1985), "A Smooth Nonparametric Estimator of a Quantile Function," Journal of the American Statistical Association, 80, 1004-1011.

END

DATE

FILMED

JAN

1988